Air pollution sources apportionment in a french urban site

Marie Chavent, Hervé Guégan, Vanessa Kuentz, Brigitte Patouille, Jérôme Saracco

Air pollution sources apportionment in a French urban site

Marie Chavent[*], Hervé Guégan[**], Vanessa Kuentz[*], Brigitte Patouille[*], Jérôme Saracco[***]

[*]IMB, UMR CNRS 5251, Université Bordeaux 1,
351 cours de la libération, 33405 Talence Cedex, France

[**]ARCANE-CENBG,
Le Haut Vigneau, BP120, 33175 Gradignan Cedex, France

[***]GREThA, UMR CNRS 5113, Université Montesquieu - Bordeaux IV
Avenue Léon Duguit, 33608 Pessac Cedex, France

Abstract

The development of air quality control strategies is a wide preoccupation for human health. In order to achieve this purpose, air pollution sources have to be accurately identified and quantified. This case study is part of a scientific project initiated by the French ministry of Ecology and Sustainable Development. Measurements of chemical composition data for particles have been realized on a French urban site. The work presented in this paper splits into two main steps. In the first one, the identification of the sources profiles has been reached thanks to Principal Component Analysis (PCA), followed by a rotation technique. Then, in the second step, a receptor modelling approach (using Positive Matrix Factorization as estimation method) allows to evaluate the apportionment of the sources. The results from these two statistical methods have enabled to characterize and apportion five sources of fine particulate emission.

Key-words: air pollution data, Principal Component Analysis (PCA), Positive Matrix Factorization (PMF), rotation.

1. Introduction

Air pollution is a complex mixture of extremely small particles and liquid droplets suspended in the air we breathe. Various sources as factory and utility smokestacks, vehicle exhaust, wood burning, mining, construction activity and agriculture, are known to generate particulate pollution, also called particulate matter (PM). High concentrations of particles have been found to present a serious danger to human health (Pope et al., 2002; Samet et al., 2000).

In this study, particles of special preoccupation to the protection of lung health are those known as fine particles, less than 2.5 microns in diameter and called PM2.5 in the rest of this paper. Thus, a wide preoccupation of environmental protection agencies concerns the development of PM2.5 control strategies. One of the main goals of these strategies is to improve ambient air quality. In consequence, this involves the reduction of emissions from primary sources. Therefore it is important to be able to identify these air pollution sources and apportion the contributions of these sources.

A reliable way to provide information regarding source characteristics is often obtained from a receptor modelling approach, using measurements of chemical composition data for particles at a sample site; see Hopke (1991) for some details and useful references. Most of the multivariate receptor models are based on the analysis of the correlations between measured concentrations of chemical species, assuming that highly correlated compounds come from the same source. Principal Component Analysis (PCA) is commonly used as multivariate receptor model, and this multivariate method has been successfully applied to identify sources in several studies.

However, PCA fails to quantify sources contributions. Then specific methods are needed to address this problem. One of them is Positive Matrix Factorization (PMF); see for instance Paatero and Tapper (1994).

The case study presented here corresponds to the statistical part of the scientific program PRIMEQUAL (Projet de Recherche Interorganisme pour une MEilleure QUalité de l'Air à l'échelle Locale), initiated by the French ministry of Ecology and Sustainable Development and the ADEME (Agence de l'Environnement et de la Maîtrise de l'Energie, that is French Environment and Energy Management Agency), about atmospheric pollution and its impacts. In this statistical work, a methodology for determining particulate emission sources and their concentrations at the urban site of Anglet located in the South-West of France has been proposed and applied. This paper is based on a more complete work presented in Chavent et al. 2007.
The following three steps process has been implemented:

(i) The air pollution data (that is PM2.5) were collected with sequential fine particle samplers on the receptor site and the chemical composition of each sample was measured with PIXE (Particle Induced X-ray Emission) method. A **data matrix of chemical compounds concentrations** in each sample has been obtained after several pre-treatments.

(ii) To **identify possible air pollution sources,** we have implemented PCA approach to this data matrix and we have rotated the standardized principal components in order to obtain more interpretable results.

(iii) For the **sources apportionment issue**, we have applied PMF to the same data matrix and normalized the results so as to find components with physical interpretations (concentration of each source in each sample).

It is interesting to note that steps (ii) and (iii) are numerically and computationally independent. Because the PMF method can be used for both identifying and quantifying the pollution sources, case studies usually don't mix PCA and PMF. But in practice, it can be difficult when using PMF to identify potential sources without some sort of profile to which to compare the final results. In this case study, we mix PCA and PMF in the sense that we check that each source quantified with PMF is clearly correlated with a single source identified with PCA. Sources difficult to identify with PMF are then clearly identified with the help of PCA and the identity of the others are validated with PCA.
The rest of this case study is organized as follows. A description of the air pollution data set is given in Section 2. The sources identification step via PCA and Varimax rotation is developed in Section 3. Section 4 is devoted to the sources apportionment step via PMF. Finally, Section 5 gives conclusion and summarizes the paper.

2. The data set

PM2.5 samplers were collected by AIRAQ[1] during December 2005 and July 2006 at the French site of Anglet located in the South-West of France (see Figure 1). In this case study, we only exhibit the results corresponding to the winter data set.

[1] Réseau de surveillance de la qualité de l'air en Aquitaine

Figure 1 : The French urban site

This sampling site located at "Station fixe d'Anglet" (see the map given in Figure 1) was chosen because of its proximity with:

- a high traffic road in red on the map,
- three cities (Bayonne, Anglet, Biarritz) with a total of 170.000 inhabitants,
- an industrial area with a steelworks and a refinery at North-East,
- Atlantic Ocean at West.

This receptor site is thus subject to different pollution origins: traffic road, urban and industrial activities, and natural dust. The knowledge of potential origins has been determinant in the choice of the site. Indeed, it allowed to check if the sources identified with the statistical methodology (without using any information about the sources) were coherent with the expected ones. The n = 61 samples of PM2.5 were collected every twelve hours: one for the day (7AM:19PM) and one for the night (19AM:7PM).

The mass, the volume and the concentration C in ng/m^3 of each particle sampler were measured with the PIXE method by ARCANE-CENBG[2], as well as the concentrations of p = 15 chemical elements (Al, Si, P, S , Cl, K, Ca, Ti, Mn, Fe, Ni, Cu, Zn, Br, Pb)[3]. Table 1 gives a subset of the data in their initial form. We notice on this data table that the concentrations of the 15 elements measured with PIXE are very small comparing to the total concentrations C in the samples. Obviously the sum of the 15 concentrations measured with PIXE doesn't add up to the total concentration of the samples. The elements (H, C, N, O)[4] not measured with PIXE represent almost all the remaining concentration.

---

[2] Atelier Régional de Caractérisation par Analyse Nucléaire Elémentaire – Centre d'Etudes Nucléaires de Bordeaux Gradignan

[3] Aluminum (Al), Silicon (Si), Phosphorus (P), Sulphur (S), Chlorine (Cl), Potassium (K), Calcium (Ca), Titanium (Ti) Manganese (Mn), Iron (Fe), Nickel (Ni), Copper (Cu), Zinc (Zn), Bromine (Br), Lead (Pb).

[4] Hydrogen (H), Carbon (C), Nitrogen (N), Oxygen (O).

Date	*C*	*Al*	*Si*	…	*K*	*Ca*	…	*Br*	*Pb*
23-11-05 day	7264.2	92	75	…	163	35	…	7	10
23-11-05 night	9633.0	135	90	…	211	23	…	7	77
24-11-05 day	10952.4	175	137	…	241	69	…	8	19
24-11-05 night	5333.3	36	31	…	94	44	…	9	7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24-12-05 day	20978.3	<2	<1	…	266	<1	…	7	18
24-12-05 night	18130.8	18	<1	…	307	<1	…	7	19
25-12-05 day	23297.9	37	22	…	311	12	…	7	14
25-12-05 night	36105.3	<2	<1	…	277	<1	…	10	19

Table 1: Subset of the original data table

Elements *Ni* and *Ti* that were frequently present at concentrations below the detection limits (BDL) were excluded and only 13 elements were selected. Then the few BDL data remaining in this data set were replaced by values corresponding to one-half the appropriate analytical detection limit. In nature, the elements *Al*, *Si*, *S* and *Fe* are usually found in the following oxidized forms: Al_2O_3, SiO_2, SO_4, Fe_2O_3. For this chemical reason, *Al*, *Si*, *S* and *Fe* were replaced by the compounds Al_2O_3, SiO_2, SO_4, Fe_2O_3: we added the mass of the measured element to the mass of oxygen of its oxidized form. Then, the remaining concentration, called C_{org}, which was not measured by the previous compounds and elements was calculated for each particle sampler:

$$C_{org} = C - (Al_2O_3 + SiO_2 + P + SO_4 + Cl + K + Ca + Mn + Fe_2O_3 + Cu + Zn + Br + Pb).$$

The addition of the column C_{org} in the data matrix is a key point specific to this case study. We have indeed noticed during the data pre-treatment that the remaining concentrations not measured by PIXE could be an important part of the samples concentrations. We can see Figure 2 that the proportion of C_{org} in the samples which have a concentration greater than 5 $\mu g/m^3$, is at least 50% of the total concentration. C_{org} will then be used to distinguish among the sources identified, those mostly participating to the concentration in PM2.5. The other 15 columns will be used to identify the sources. For instance, it is known that the elements *Zn* and *Pb* are in particulates emitted by industrial sources.

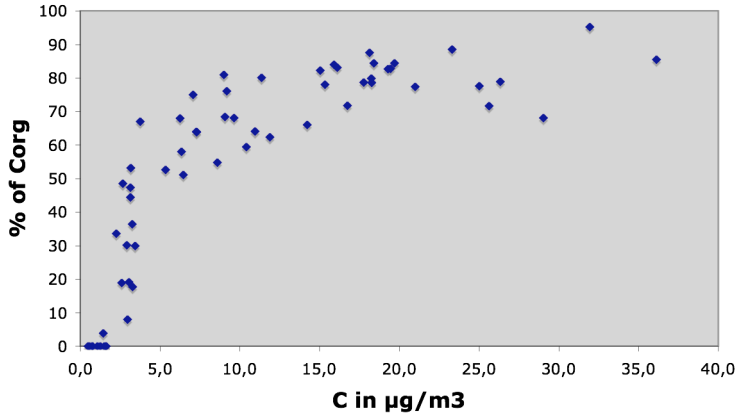

Figure 2: The proportion of C_{org} in the sample according to the total concentration *C* of the sample

Table 2 shows the data set after having applied the transformations described above. The (n,p) concentration matrix $X = (x_{ij})$ used in the receptor model has then $n = 61$ rows and $p = 14$

columns (Al_2O_3, $Si0_2$, P, $S0_4$, Cl, K, Ca, Mn, Fe_20_3, Cu, Zn, Br, Pb, C_{org}). The coefficient x_{ij} is the concentration of the *j*th chemical compound in the *i*th sample.

Date	C	Al_2O_3	$Si0_2$	…	K	Ca	…	Br	Pb	C_{org}
23-11-05 day	7264.2	250	160	…	163	35	…	7	10	4645.3
23-11-05 night	9633.0	365	193	…	211	23	…	7	77	6564.2
24-11-05 day	10952.4	475	292	…	241	69	…	8	19	7017.1
24-11-05 night	5333.3	96	66	…	94	44	…	9	7	2805.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24-12-05 day	20978.3	3	1	…	266	0.5	…	7	18	16240.7
24-12-05 night	18130.8	49	1	…	307	0.5	…	7	19	15878.3
25-12-05 day	23297.9	101	46	…	311	12	…	7	14	20636.8
25-12-05 night	36105.3	3	1	…	277	0.5	…	10	19	30885.1

Table 2: The final data set

A meteorological data set was also used to interpret and validate some results. Hourly temperatures and wind directions (in degrees) were collected during the sampling period at a meteorological station located 2.5 *km* far from the sampling site. Temperatures were averaged to match with the 12-hours samples. The 360° of the wind directions were split into 8 categorical wind directions (North, North-East, North-West, South….) and a wind direction data matrix of 61 rows (12-hours samples) and 8 columns (wind directions) were constructed. An element of this matrix is the percentage of hours during which the wind direction has been observed (see Table 3).

Date	N	N-E	E	S-E	S	S-W	W	N-W
23-11-05 day	17%	0%	25%	8%	8%	0%	0%	42%
23-11-05 night	25%	17%	33%	25%	0%	0%	0%	0%
24-11-05 day	0%	0%	8%	17%	50%	25%	0%	0%
24-11-05 night	0%	0%	0%	0%	8%	9%	58%	25%
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 3: Wind directions data

3. PCA and Varimax rotation for sources identification

In order to identify the sources of fine particulate emission, we applied PCA to the concentration matrix *X*. The idea was to find groups of correlated chemical compounds that are characteristics of air pollution sources. For instance, if the elements *Zn* and *Pb* are strongly correlated to the same factor, since these elements are known to have industrial origin, this factor will be associated to industrial pollution source.

In PCA, we consider a (*n*,*p*) numerical data matrix *X* where *n* objects are described on $p < n$ variables $x_1,\ldots,x_p$. We note x_j a column of *X*. Let $\tilde{X} = \left(\tilde{x}_{ij}\right)_{n,p}$ be the standardized data matrix with $\tilde{x}_{ij} = \dfrac{x_{ij} - \bar{x}_j}{s_j}$ where $\bar{x}_j$ and s_j are the sample mean and the sample standard deviation of x_j. The basic idea underlying Factor Analysis (using correlation matrix) is that the *p* observed

standardized variables $\tilde{x}_1,\ldots,\tilde{x}_p$, can be expressed, except for an error term, as linear functions of $q < p$ unobserved variables or common factors $f_1,\ldots,f_q$. Given the observed standardized matrix $\tilde{X}$, Factor Analysis model can be expressed in its simplified form as:

$$\tilde{X} = FA' + E,$$

where F is the (n,q) matrix of unobserved values of the factors, whereas the (p,q) matrix A is the unknown loading matrix which provides information that relates the factors $f_1,\ldots,f_q$ to the original variables $x_1,\ldots,x_p$. Several approaches were developed to estimate the model (principal factor, maximum likelihood, …) but PCA is often used in practice. With PCA estimation method, the q columns of F are the first q standardized principal components (which are mutually orthogonal and of variance equal to 1) and each element a_{jk} of A is equal to the correlation between the variable x_j and the factor f_k.

In this case study, we applied PCA to the concentration matrix X where the 61 samples (in row) are described by 14 compounds (in column). We see Figure 3 that each k^{th} column of F and k^{th} row of A' obtained with PCA will be associated to a source. The approach is the following: we search in the k^{th} row of the loading matrix A', the compounds strongly correlated with the k^{th} factor. If these compounds are known to be characteritisc from a source, this source is associated to this factor. Because the n samples are chronologically ordered, the k^{th} column of the factor score matrix F gives an idea of the evolution of the quantity of fine particulate emitted by the source associated with the k^{th} factor.

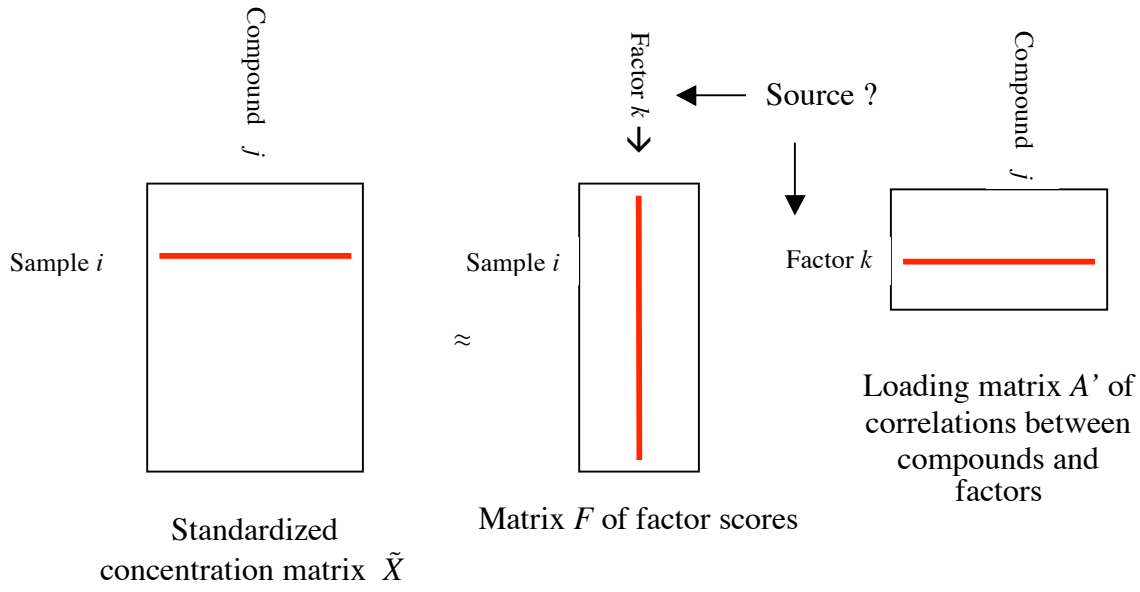

Figure 3: Decomposition of the standardized concentration matrix

Of course, depending of the correlations in the loading matrix, it is not always possible to associate clearly a source to a factor. In this study, after several trials, we have chosen q=5. Indeed, it was not possible to associate clearly a source to each factor with the decompositions in more than 5 factors. Moreover, with q=5, we have 90,93% of total variance that is explained.

Table 4 gives the loading matrix *A'* obtained with the following procedure FACTOR of SAS:

```
PROC FACTOR data=hiverorg method=prin nfactors=5 outstat=load;
var Al2O3 SiO2 P SO4 Cl K Ca Mn Fe2O3 Cu Zn Br Pb Corg;
run;
```

where:

- data=hiverorg is the sas dataset constructed from the concentration data matrix (Table 2)
- method=prin because the factor analysis method is PCA
- outstat=load is the sas dataset with the loadings reported Table 4

	f_1	f_2	f_3	f_4	f_5
Al_2O_3	0,672	-0,663	0,221	-0,193	0,023
SiO_2	0,649	-0,669	0,254	-0,203	-0,058
P	0,682	-0,629	0,223	-0,238	0,006
SO_4	0,589	0,449	-0,413	-0,220	0,175
Cl	-0,474	-0,281	0,192	0,681	0,344
K	0,888	-0,154	-0,209	0,004	0,284
Ca	0,638	-0,405	0,099	0,399	-0,195
Mn	0,384	0,776	0,183	0,093	-0,247
Fe_2O_3	0,793	0,319	0,038	0,269	-0,360
Cu	0,796	0,248	-0,098	0,232	-0,360
Zn	0,352	0,663	0,589	-0,072	0,246
Br	0,746	-0,182	-0,126	0,352	0,363
Pb	0,428	0,659	0,519	-0,088	0,296
C_{org}	0,600	0,297	-0,613	-0,025	0,222

Table 4: Correlations between the chemical compounds and the 5 factors

Since most compounds are positively correlated with the first standardized principal component, it is difficult to detect groups of correlated elements.

However the graphical representation of these compounds according to their correlations with f_1 and f_2 (see Figure 4) or according to their correlations with f_2 and f_3 (see Figure 5) shows groups of compounds that seem to be correlated with each other (*Zn* and *Pb* for instance or *P*, Al_2O_3 and SiO_2). Because it is known that *Zn* and *Pb* for instance come from fine particulates with industrial origin, we would like to see clear correlations between those two elements and a factor.

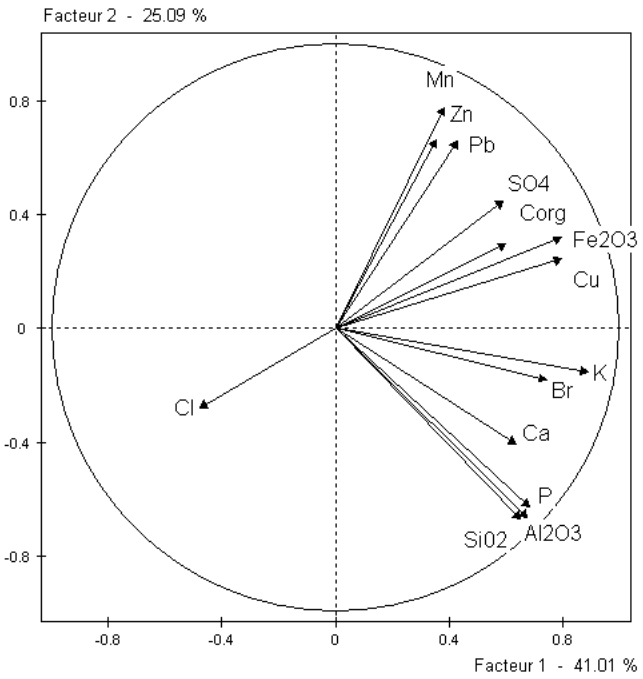

Figure 4: Factor 1-2 correlation circle

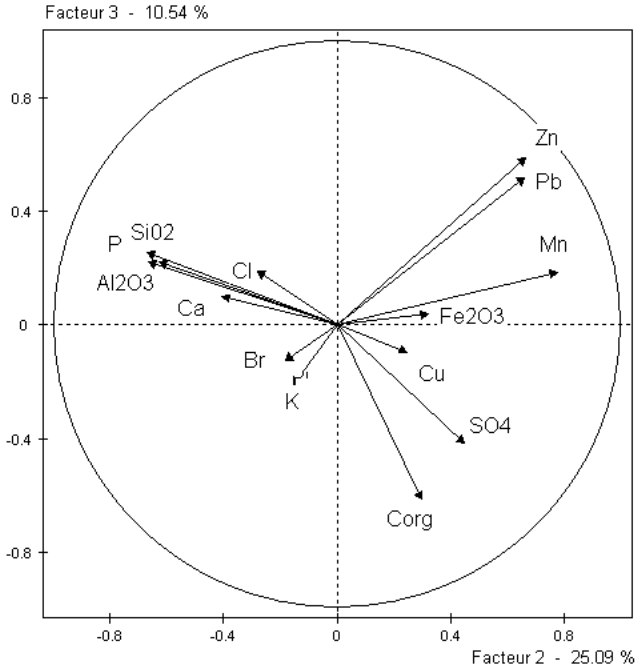

Figure 5: Factor 2-3 correlation circle

In order to identify more clearly groups of correlated compounds, a Varimax rotation has been applied to the standardized principal components. The idea of rotation in Factor Analysis is the

following. Let T be an orthogonal transformation matrix, $TT' = T'T = I_q$. The factor analysis model can then be expressed as:

$$\tilde{X} = \underbrace{FT}_{\tilde{F}}\underbrace{T'A'}_{\tilde{A}} + E$$

with:

- $\tilde{F} = FT$ the matrix of the rotated standardized principal components (rotated factors) which are still mutually orthogonal and of variance equal to 1,
- $\tilde{A} = AT$ the matrix of rotated loadings which are correlations between the variables and the rotated principal components.

From a practical point of view, the orthogonal transformation matrix T is defined in order to construct a matrix $\tilde{A}$ such that each variable x_j is clearly correlated to one of the rotated factor $\tilde{f}_k$ (that is $\tilde{a}_{jk}$ close to 1) and not to the other rotated factors (that is $\tilde{a}_{jk^*}$ close to 0 for $k^* \neq k$).

The most popular rotation technique is Varimax. It seeks rotated loadings that maximize the variance of the squared loadings in each column of $\tilde{A}$.

The matrix $\tilde{A}$ of the loadings after rotation and the matrix $\tilde{F}$ of the rotated factor scores are obtained with the procedure FACTOR of SAS:

```
PROC FACTOR data=hiverorg method=prin nfactors=5 outstat=load out=fact
rotate=varimax;
var Al2O3 SiO2 P SO4 Cl K Ca Mn Fe2O3 Cu Zn Br Pb Corg;
run;
```

where:

- outstat=load is the sas dataset containing the rotated loading
- out=fact is the sas dataset containing the rotated factor scores

	$\tilde{f}_1$	$\tilde{f}_2$	$\tilde{f}_3$	$\tilde{f}_4$	$\tilde{f}_5$
Al_2O_3	0,981	0,087	-0,042	0,070	-0,038
SiO_2	0,979	0,012	-0,055	0,104	-0,074
P	0,972	0,090	-0,017	0,071	-0,092
SO_4	-0,028	0,765	0,247	0,180	-0,345
Cl	-0,153	-0,274	-0,136	-0,181	0,879
K	0,597	0,716	0,111	0,233	0,031
Ca	0,608	0,091	-0,113	0,560	0,272
Mn	-0,279	0,119	0,604	0,582	-0,238
Fe_2O_3	0,198	0,282	0,289	0,848	-0,112
Cu	0,213	0,359	0,161	0,816	-0,149
Zn	-0,029	0,053	0,977	0,129	-0,044
Br	0,490	0,615	0,097	0,281	0,392
Pb	0,004	0,163	0,969	0,126	-0,054
C_{org}	-0,018	0,893	0,021	0,222	-0,160

Table 5: Correlations between the chemical compounds and the 5 rotated factors

The loadings after rotation are reported Table 5. These new loadings are correlations between the rotated factors and the compounds. We notice now that the five rotated factor $\tilde{f}_1$, ..., $\tilde{f}_5$ are clearly correlated with some compounds (highlighted in grey in the table). Because some of these compounds are known to be characteristic from specific pollution sources, we have il is

possible here to associate five sources to the 5 rotated factors: $\tilde{f}_5$ is clearly correlated with *Cl*, which is known to have sea salt origin. In the same way, the elements *Zn* and *Pb* correlated with $\tilde{f}_3$ have industrial origin and the compounds Al_2O_3 and $Si0_2$ correlated with $\tilde{f}_3$ are coming from soil dust. Factors $\tilde{f}_1$, $\tilde{f}_3$ and $\tilde{f}_5$ can then be clearly associated to those three sources (see Table 6).

Rotated factor	Source	Characteristic correlated elements or compounds
$\tilde{f}_1$	Soil dust	Al_2O_3, $Si0_2$
$\tilde{f}_2$	Combustion	$S0_4$
$\tilde{f}_3$	Industry	*Zn*, *Pb*
$\tilde{f}_4$	Vehicle	Fe_20_3, *Cu*
$\tilde{f}_5$	Sea	*Cl*

Table 6: Factor-source associations

In the same way, $S0_4$ correlated with $\tilde{f}_2$ is usually linked to combustion and Fe_20_3 and *Cu* correlated with $\tilde{f}_4$ can be linked linked traffic road. In order to confirm these two last associations we have confronted the rotated factors $\tilde{f}_k$ with external informations such as meteorological data (temperatures and wind directions) and the periodicity night/day of the sampling. The coefficients $\tilde{f}_{ik}$ of the column represent here a "relative" contribution of the source *k* to the sample *i*.

Figure 6 gives the evolution of the relative contribution of the source associated with $\tilde{f}_4$. The night samples have been distinguished from the day ones, which enables to notice that the contribution of this source is stronger during the day than at night. It is then a confirmation that this source corresponds to vehicle pollution.

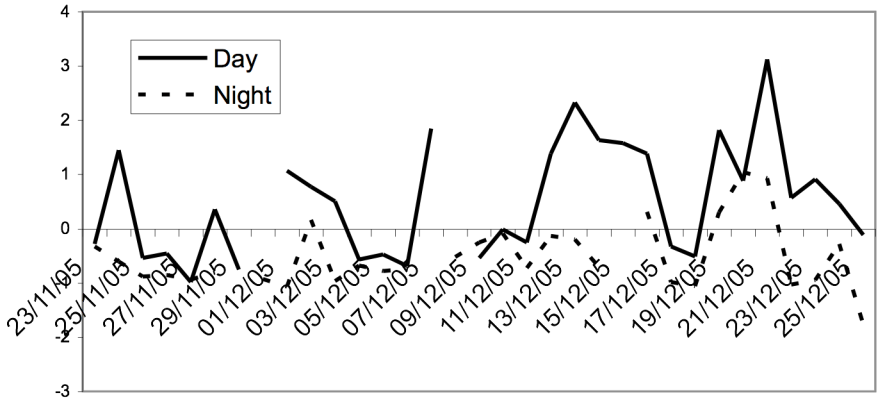

Figure 6: Evolution of the rotated factor 4 associated to cars pollution

Figure 7 gives the evolution of the relative contribution of the source associated with $\tilde{f}_2$. We notice an increase in the contribution of this source at the middle of the sampling period, which corresponds to a decrease in the temperature measured on the sampling site, see Figure 8.

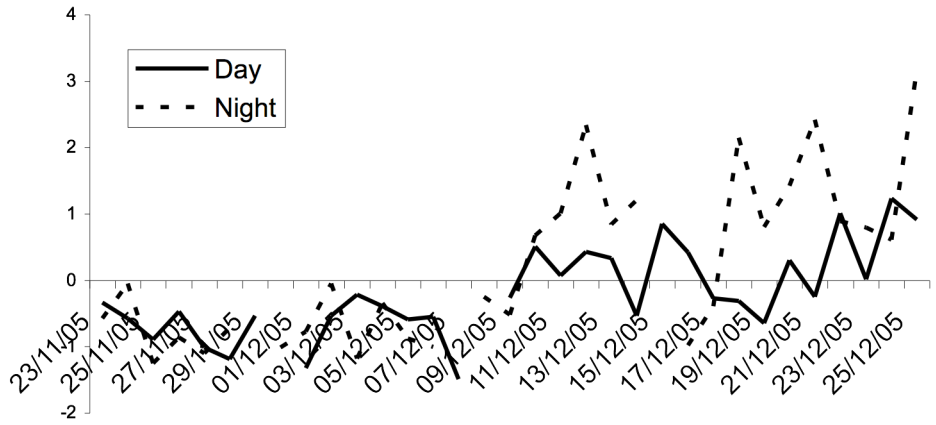

Figure 7: Evolution of the rotated factor 2 associated to heating source

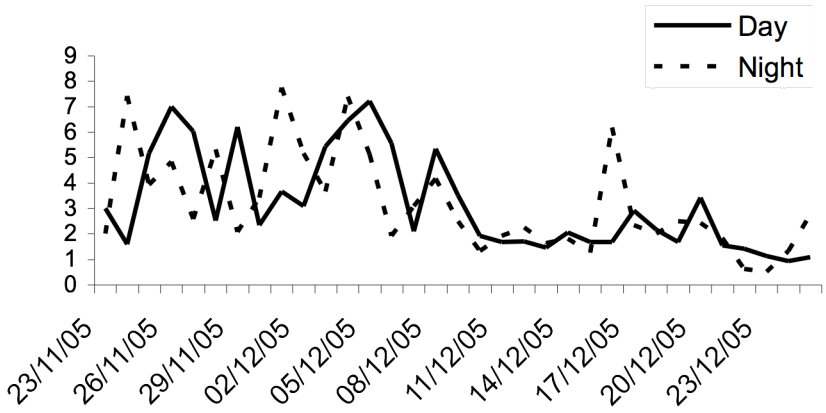

Figure 8: Evolution of temperatures

Finally, correlations between the rotated factor scores and the eight vectors of wind directions (see Table 3) have been calculated. Figure 9 (resp. Figure 10) is a graphical representation of the correlations between $\tilde{f}_3$ (resp $\tilde{f}_5$) and wind directions.

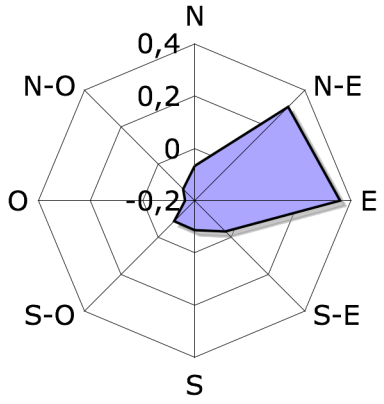

Figure 9: Correlations between the rotated factor 3 associated to industrial source and the 8 wind directions

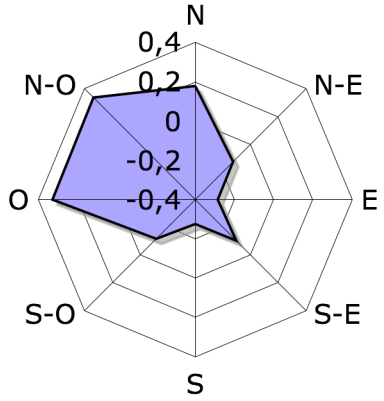

Figure 10: Correlations between the rotated factor 5 associated to sea source and the 8 wind directions

The correlation of $\tilde{f}_3$ with East and North-East wind directions confirms that this source corresponds to industrial pollution. Indeed, the main industries are located at North of the sampling site (see Figure 1). In the same way, the correlation of $\tilde{f}_5$ with West and North-West wind directions is a confirmation that this source corresponds to Atlantic Ocean located at West of the sampling site, and then to sea pollution.

4. Sources apportionment via PMF

We have seen how fine particulate pollution sources can be identified by applying PCA to a concentration data matrix. But the identification is not sufficient. The danger for the health is indeed linked to the quantity of PM2.5 we breath. The problem is then not only to identify the PM2.5 sources but also to determine in which proportion these sources participate to the global dust contamination. In order to quantify unknown sources of fine particulate emission, we have

approximated a receptor model by first defining a Positive Matrix Factorisation of the concentration matrix X, and then by normalizing the results to find components with physical interpretations.

The basic problem of receptor modelling is to estimate, from the data matrix X and the number q of sources, their compositions and their contributions. To address this problem, we consider the mass balance equation:

$$x_{ij} = \sum_{k=1}^{q} g_{ik} b_{jk}$$

where

- x_{ij} is the concentration of the jth chemical species in the ith sample,
- g_{ik} is the concentration in particles from source k in the sample I,
- b_{jk} is the mass fraction (percentage) of species j in source k.

In the common parlance of receptor modelling, the b_{jk}'s are the sources compositions (or sources profiles) and the g_{ik}'s are the sources contributions. The product $g_{ik}b_{jk}$ is then the approximation of the concentration in the sample i in particles from the jth species coming from the source k. Let m_{ijk} be the mass, in the sample i, of species j from source k, and let m_{ik} be the mass in the sample i from source k. Then $b_{jk} = \frac{m_{ijk}}{m_{ik}}$ is a mass fraction or, in other words, it is the percentage of species j emitted by source k when sample i was collected. Since the mass fraction b_{jk} is independent from i, it means that the sources profiles are assumed to be constant during the sampling period.

In matrix form, the mass balance equation can be written:

$$X = GB'$$

where G is a (n,q) matrix of sources contributions and B is a (p,q) matrix of sources compositions (see Figure 11).

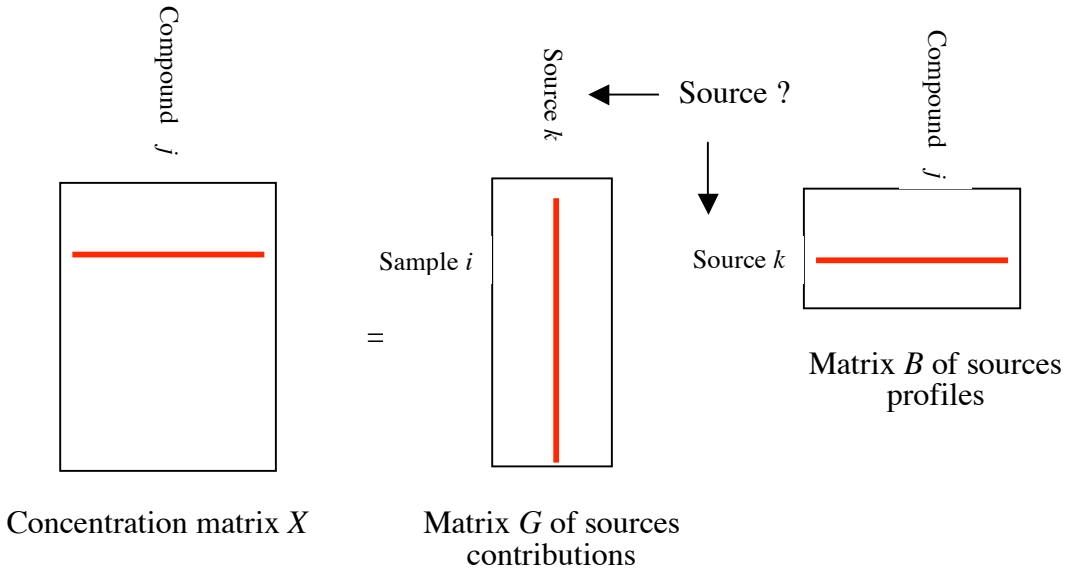

Figure 11: Decomposition of the concentration matrix

In this case study, 5 sources were clearly identified with PCA. We have then chosen to approximate the matrices G and B for $q = 5$ in the receptor model. Two steps (PMF and normalization) have been necessary to approximate G and B. Once these approximations $\hat{G}$ and $\hat{B}$ calculated, the user has an approximation of the quantities of fine particulate emitted by 5 sources in each sample. He has also an approximation of the profiles of the sources. But no name is associated to each source. As we will a supplementary step is necessary to identify clearly the sources quantified in $\hat{G}$.

PMF step. The matrix X is factorized in a product HC' of rank q under constraints of positivity of the coefficients. This condition is required by physical reality of non-negativity of sources compositions and contributions: $g_{ik} \geq 0$ and $b_{jk} \geq 0$.

The PMF algorithm developed by Paatero and Tapper (1994) in the context of receptor modelling minimizes

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\frac{x_{ij} - \sum_{k=1}^{q} h_{ik}c_{jk}}{\sigma_{ij}}\right)^2$$

subject to $h_{ik} \geq 0$ and $c_{jk} \geq 0$. The coefficient σ_{ij} is a measure of uncertainty of the observation x_{ij}. In this case study, we are dealing with variables measured on very different scales which can cause problems when approximating X globally on all the variables (for instance, minimizing an unweighed quadratic error, that is when $\sigma_{ij} = 1$, will give better approximations for the columns of X corresponding to variables with large dispersion). Hence we have opted for $\sigma_{ij} = s_j$, the sample standard deviation of the jth variable.

The numerical results have been obtained with the PMF algorithm and program poposed by Jianhang et Laosheng (2004). This program takes as input:

- the concentration matrix $X_{61\times14}$,
- the matrix of the uncertainty measures $(\sigma_{ij} = s_j)_{61\times14}$
- the number of sources q=5

It gives in output a matrix $\hat{H}_{61\times5}$ and a matrix $\hat{C}_{14\times5}$, solution of the above constrained minimization problem. Obviously, this solution is not unique and other physical constraints have been used to calculate from these two matrices $\hat{H}_{61\times5}$ and $\hat{C}_{14\times5}$ the approximations $\hat{G}$ and $\hat{B}$ of the contribution matrix and of the profile matrix. This is the scaling step. Let $\hat{X} = \hat{H}\hat{C}'$ be the product calculated by PMF.

Scaling step. The columns of the approximations $\hat{H}$ and $\hat{C}$ obtained in the previous step, must be scaled in order to get the approximations $\hat{G}$ and $\hat{B}$. The scaling coefficients are defined to fulfil other physical constraints of the sources compositions and contributions.

Let us first remark that $\hat{x}_{ij} = \sum_{k=1}^{q} \hat{h}_{ik}\hat{c}_{jk} = \sum_{k=1}^{q} \hat{h}_{ik}\frac{\beta_k}{\beta_k}\hat{c}_{jk}$. Then, the matrix $\hat{X}$ can be written: $\hat{X} = \breve{H}\breve{C}'$ with $\breve{h}_{ik} = \hat{h}_{ik}\beta_k$ and $\breve{c}_{jk} = \frac{\hat{c}_{jk}}{\beta_k}$. The objective of scaling is then to define the scaling constants β_k, $k = 1,\ldots,q$ such that $\breve{H}$ and $\breve{C}$ verify the physical conditions of the matrices *G* and *B* of the mass balance equation. In this study, we consider the two following conditions:

(a) The sum of the concentrations of the sources adds up to the total concentration of the samples; that is for each sample, $\gamma_i = \sum_{k=1}^{q} g_{ik}$ where γ_i is the total concentration in the *i*th sample.

(b) If the sum of the concentrations of the observed species adds up to the total concentration of the samples, then the sum of all species in each source profile is equal to unity, that is:

$$\sum_{j=1}^{p} b_{jk} = 1 \text{ if } \sum_{j=1}^{p} x_{ij} = \gamma_i .$$

Note that the introduction in this case study of the 15th column C_{org} (the concentration in the samples not measured by PIXE) calculated for each particle sampler yields $\sum_{j=1}^{p} x_{ij} = \gamma_i$.

From the physical constraint (b), the scaling coefficients β_k can be directly calculated: from $\hat{X} = \hat{H}\hat{C}'$, we get $\hat{\beta}_k = \sum_{j=1}^{p} \hat{c}_{jk}$. Details can be found in Chavent et al. (2007).

In practice, we have imported in Excel the numerical values of the matrices $\hat{H}_{61\times5}$ and $\hat{C}_{14\times5}$ obtained in output of the PMF program and calculated:

- the 5 scaling coefficients (given Table 7) by summing the 5 columns of the matrix $\hat{C}_{14\times5}$,
- the (14,5) matrix $\hat{B}$ of the approximated compositions (profiles) of the 5 sources on the 14 compounds, by dividing the five columns of $\hat{C}_{14\times5}$ by the corresponding scaling coefficient,
- the (61,5) matrix $\hat{G}$ of the approximated concentrations of the 5 sources in the 61 samples, by multiplying the five columns of $\hat{H}_{61\times5}$ by the corresponding scaling coefficient.

k	$\hat{\beta}_k$
1	147.1
2	91.5
3	73.5
4	251.9
5	51.1

Table 7: The scaling coefficients

Quality of the model approximation. We can evaluate the quality of the approximation of X by $\hat{G}\hat{B}'$: the Figure 12 shows a good fitting of the γ_i's by the $\hat{\gamma}_i$'s with $\hat{\gamma}_i = \sum_{k=1}^{q} \hat{g}_{ik}$.

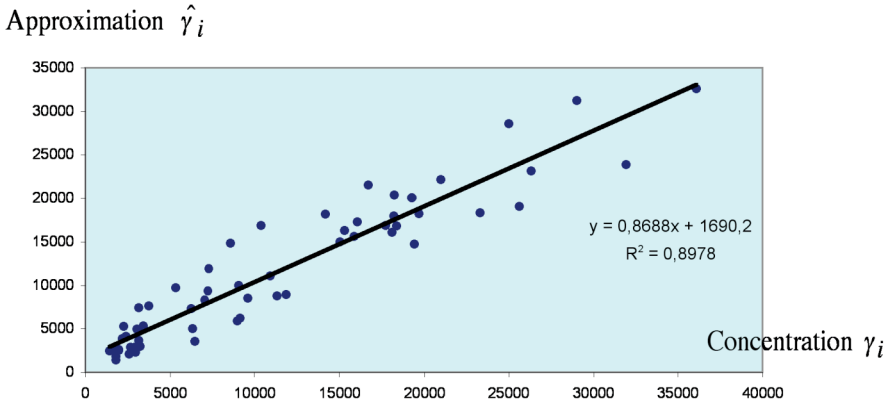

Figure 12: Adjustment of γ by $\hat{\gamma}$

Sources identification. In practice, the knowledge of $\hat{G}$ and $\hat{B}$ gives no direct indications on the nature of the sources. To try to discover the nature of the five sources, we want to calculate their relative contribution to each of the 14 chemical compounds. To do that, we need to work with the masses instead of the concentrations. Then we calculate, from $\hat{G}$, the approximation of the total mass of particulate emitted from source k in the 61 samples. This mass is multiplied by $\hat{b}_{jk}$ and it gives the percentages reported in Table 8.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Al_2O_3	100.0	0.0	0.0	0.0	0.0
$Si0_2$	100.0	0.0	0.0	0.0	0.0
P	81.5	0.5	3.9	8.2	6
$S0_4$	4.5	9.5	10.7	67.9	7.5
Cl	0.0	0.0	0.0	0.0	100.0
K	38.8	0.0	4.4	56.7	0.2
Ca	42.0	39.6	0.0	0.0	18.4
Mn	0.0	54.9	33.1	8.5	3.5
Fe_20_3	19.0	59.2	14.4	7.4	0.0
Cu	18.5	56.8	9.1	15.6	0.0
Zn	9.0	0.5	87.5	0.0	3.1
Br	19.4	12.1	5.7	33.4	29.4
Pb	10.7	0.0	81.4	7.9	0.0
C_{org}	0.0	8.0	0.0	92.0	0.0

Table 8: Relative contributions of the sources to the chemical compounds

Table 8 is used to identify the nature of the sources. For instance, Al_20_3 and SiO_2 are exclusively emitted by source 1. Because Al_20_3 and SiO_2 are known to have natural origin, this source is associated to the soil dust pollution source. We proceed in the same way for the other sources. We deduce possible identifications of the five pollution sources, see Table 9.

$k = 1$	Soil dust
$k = 2$	Vehicles
$k = 3$	Industry
$k = 4$	Combustion
$k = 5$	Sea

Table 9: Receptor model sources identification

One can remark that the sources identified in Table 9 are the same than those found with PCA in Table 6. To verify the coherence of these sources identifications, we have calculated, in Table 10, the correlations between the factors obtained with PCA and the sources obtained by receptor modelling (the columns of $\hat{G}$). We observe that the factors match well with the receptor model sources.

	Source 1	Source 2	Source 3	Source 4	Source 5
Factor 1	**0.98**	-0.18	-0.11	-0.02	-0.18
Factor 2	0.11	0.12	0.06	**0.95**	-0.30
Factor 3	-0.05	-0.09	**0.98**	0.02	-0.15
Factor 4	0.12	**0.96**	0.10	0.11	-0.22
Factor 5	-0.02	-0.13	-0.10	-0.27	**0.88**

Table 10: Correlations between the sources of the receptor model and the factors of PCA

Sources descriptions. The matrix $\hat{B}$ of the sources profiles is reported in Table 11. We notice that, according to these profiles, C_{org} which represents almost the total concentration in PM2.5, is only emitted by the vehicle and combustion sources. It is also possible to describe the composition of the sources using Table 11: for instance, sea pollution source is made of around 75% of Chlorine and 23% of $S0_4$.

	Soil dust	Vehicles	Industry	Combustion	Sea
Al_2O_3	41.6	0.0	0.0	0.0	0.0
$Si0_2$	18.5	0.0	0.0	0.0	0.0
P	6.2	0.0	0.7	0.0	0.6
$S0_4$	10.1	15.3	59.6	12.2	22.6
Cl	0.0	0.0	0.0	0.0	74.5
K	12.9	0.0	3.6	1.5	0.1
Ca	2.4	1.6	0.0	0.0	1.4
Mn	0.0	0.2	0.3	0.0	0.0
Fe_20_3	6.7	15.0	12.7	0.2	0.0
Cu	0.3	0.7	0.4	0.0	0.0
Zn	0.7	0.0	16.3	0.0	0.3
Br	0.2	0.1	0.2	0.0	0.5
Pb	0.3	0.0	6.2	0.0	0.0
C_{org}	0.0	67.1	0.0	85.9	0.0

Table 11: The sources profiles

Sources apportionments. From matrix $\hat{G}$ of the source contributions, we can deduce some interesting comments. First we can focus on the relative contribution of each source in each particle sampler. For instance, Figure 13 represents the relative contributions of the combustion

source in the 61 particle samplers. We can notice the increase in the part of this source in the second period of sampling, corresponding to a decrease in the temperature (see Figure 8).

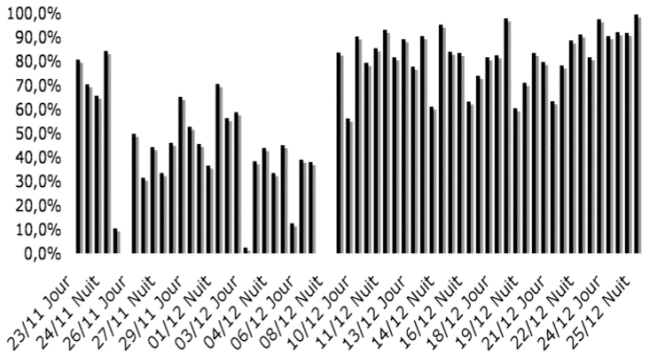

Figure 13: Relative contribution of the source combustion to the samples

We can also focus on the contribution of the sources to the PM2.5 dust contamination during the sampling period. Figure 14 shows the domination of the combustion source during this winter sampling period.

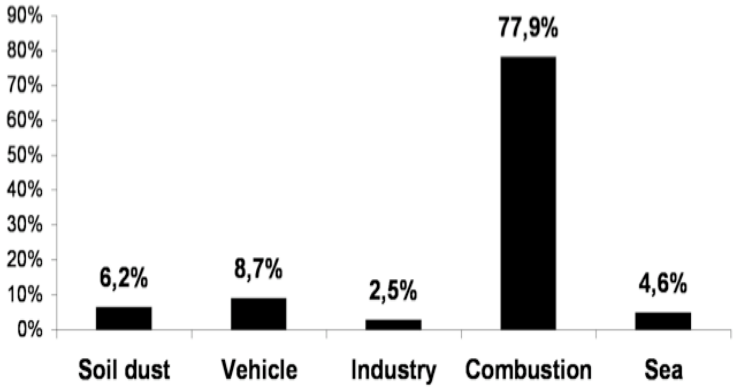

Figure 14: Global sources contributions to the PM2.5 dust contamination

5. Conclusion

We exhibited in this case study a methodology in order to determine air pollution sources apportionment in a French urban site. The data correspond to measurements of chemical composition data for particles. Our approach is divided into two parts: first we identify the sources profiles via Principal Component Analysis followed by a Varimax rotation. Then, we evaluate the apportionment of the sources via a receptor modelling approach based on Positive Matrix Factorization as estimation method. The corresponding numerical results allow to characterize and apportion five principal sources of fine particulate emission: soil dust, vehicles, industry, combustion and sea. These results are not hardly surprising, they confirm the environmental knowledge of the phenomenon of air pollution by fine particles. What is interesting here is the fact that we do not use any prior information in order to retrieve the usual

five air pollution sources. This kind of methodology can then be applied to any dataset and problems of identification and apportionment of sources. For mathematical details on the proposed methodology, the reader can refer to Chavent et al. (2007).

6. References

Hopke, P.K., 1991. *Receptor Modeling for Air Quality Management*. Elsevier, Amsterdam.

Paatero, P., Tapper, U., 1994. Positive Matrix Factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, **5**, 111-126.

Pope, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Journal of the American Medical Association* **287**, 1132–1141.

Chavent, M., Guégan, H., Kuentz, V., Patouille, B., Saracco, J., 2007. PCA and PMF based methodology for air pollution sources identification and apportionment. *Submitted paper*.

Samet, J.M., Dominici, F., Curriero, F.C., Coursac, I., Zeger, S.L., 2000. Fine particulate air pollution and mortality in 20 US cities, 1987–1994. *New England Journal of Medicine,* **343 (24)**, 1742–1749.

Jianhang, L., Laosheng, W., 2004. Technical details and programming guide for a general two wat positive matrix factorization algorithm. *Journal of Chemometrics* **18**, 519-525.